Linear Regression analysis

Silvia Saloni

Monday, July 25, 2016

Introduction

The aim of the project was to be able to find a model that best predict the relationship between: * the number of people counted with the survey for a given class at a particular hour * Wi-fi log counted in that room at that hour

This will allow to see whether Wi-Fi log is a good predictor for estimating occupancy in a classroom.

We first tried to see if the relationship between this two variables was linear. To do so we run a linear regression.

Below we describes step by step all the analysis performed.

ANALYSIS

DATABASE CONNECTION AND DATASET

```
## Loading required package: DBI
```

First of all, we set up the connection to the database, using the following code:

```
connection <- dbConnect(MySQL(),user="root", password="",dbname="who_there_db", host="localhost")
```

Then we made a query to the database, in order to get all the groundth truth data collected in room B.002, B.004 and B.006 from 9 to 17 and the correspondent Wi-Fi Log measured in that time frame and rooms.

The dataset created had in total 216 rows and it will allow us to explore if Wi-Fi log is a good predictor of the observed occupancy of the room in a certain hour.

As a **target features** for our linear regression we decided to use the number of associated client, calculated multiplying the percentage of the room full with the capacity of the room.

As response variables or feature we considered Wi-Fi logs, which were summarised either as average of the logs counted for each room and for each hour or as maximum of the logs measured for each room and for each hour.

Together with the Wi-Fi log, we included in the data set the following features:

- **Date**, which we did not use in this analysis, because they just cover 2 weeks of Novemeber, but for future analyses they can be used to group observations by seasons or semesters or to finds seasonal trends for time series analyses.
- **Time**, which will be explored either as continous variable and as categorical to explore if the time of the day can have an affect on the Wi-Fi log. To do so we, bin the time in 4 ranges: early morning (9-11), late morning (11-13), early afternoon (13-15) and late afternoon (15-17). This will allow us to see if the Wi-Fi log accuracy was changing during the day. For example, it is more likely that all the elctronic devices are fully powered early in the morning and consequently the Wi-fi log data can be more accurate or overestimating the occupancy of the room (i.e. more than one device per person). On the contrary in the afternoon, the devices may be more likely to be out of battery and it is possible that there are less devices in the room.

- **Module**, which we are not going to include in the analysis because the majority of the module present are for computer science, but for future analyses it will be possible to explore if the Wi-fi log accuracy in predicting the occupancy change acroos the courses. Science course or computer science course will more likely to use electronic devices during lecture than art students.
- **Course level**, which can indicate us whether electronic devices will be less used during different course level. For example, first and second level courses can be less practical and therefore laptop are not needed and that can decrease the number of devices connected. On the other hand, undergraduate might be more distracted during lecture and look at their phones during lectures. This will result in an increase of connection in that hour.
- **Tutorial**, which can affect the number of logged people. First of all, because tutorial divided the room in 2 and therefore there will be measured less people than expected.
- **Double_module**, categorical variable indicating wether in the class there are more than one module, increasing the number of people expected in the room.
- **Double_module**, categorical variable indicating wether in the class went ahead to check for false positive.

The resulting data set is printed below:

```
head(AnalysisTable)
```

```
##   Room       Date Time    Module Course_Level Tutorial Double_module
## 1    1 2015-11-03    9         0            0        0             0
## 2    1 2015-11-04    9 COMP30190            3        0             0
## 3    1 2015-11-05    9         0            0        0             0
## 4    1 2015-11-06    9 COMP30220            3        0             0
## 5    1 2015-11-09    9 COMP30190            3        0             0
## 6    1 2015-11-10    9         0            0        0             0
##   Class_went_ahead Capacity Percentage_room_full Wifi_Average_clients
## 1                1       90                 0.00               4.7500
## 2                1       90                 0.25              13.4545
## 3                1       90                 0.00               6.8333
## 4                1       90                 0.00               2.4167
## 5                1       90                 0.25              14.7273
## 6                1       90                 0.00               2.2727
##   Wifi_Max_clients Survey_counted_clients   Factor_Time
## 1               21                   0.0 Early Morning
## 2               15                  22.5 Early Morning
## 3               29                   0.0 Early Morning
## 4                3                   0.0 Early Morning
## 5               18                  22.5 Early Morning
## 6               14                   0.0 Early Morning
```

DATA QUALITY REPORT

Before running any analyses, we carried out the data quality report to check for any issue related to the variable (e.g. outlier, skewed distribution, NaN values) and solutions we will implement to solve them.

Initially set all the categorical variables as factors and then we printed the descriptive statistic for all the features.

```
summary(AnalysisTable)
```

```
##       Room       Date                Time          Module
##  Min.   :1   Length:216         Min.   : 9.00   Length:216
##  1st Qu.:1   Class :character   1st Qu.:10.75   Class :character
##  Median :2   Mode  :character   Median :12.50   Mode  :character
##  Mean   :2                      Mean   :12.50
##  3rd Qu.:3                      3rd Qu.:14.25
##  Max.   :3                      Max.   :16.00
##  Course_Level    Tutorial       Double_module Class_went_ahead
##  0:63         Min.   :0.00000   0:210         0: 22
##  1:14         1st Qu.:0.00000   1:  6         1:194
##  2:23         Median :0.00000
##  3:76         Mean   :0.02778
##  4:40         3rd Qu.:0.00000
##               Max.   :1.00000
##     Capacity     Percentage_room_full Wifi_Average_clients
##  Min.   : 90.0   Min.   :0.00         Min.   :  0.00
##  1st Qu.: 90.0   1st Qu.:0.00         1st Qu.: 11.61
##  Median : 90.0   Median :0.25         Median : 23.67
##  Mean   :113.3   Mean   :0.25         Mean   : 30.33
##  3rd Qu.:160.0   3rd Qu.:0.25         3rd Qu.: 43.88
##  Max.   :160.0   Max.   :1.00         Max.   :192.92
##  Wifi_Max_clients Survey_counted_clients          Factor_Time
##  Min.   :  0.00   Min.   :  0.00         Early Morning  :54
##  1st Qu.: 18.75   1st Qu.:  0.00         Late Morning   :54
##  Median : 32.50   Median : 22.50         Early Afternoon:54
##  Mean   : 40.01   Mean   : 28.01         Late Afternoon :54
##  3rd Qu.: 55.25   3rd Qu.: 40.00
##  Max.   :230.00   Max.   :160.00
```

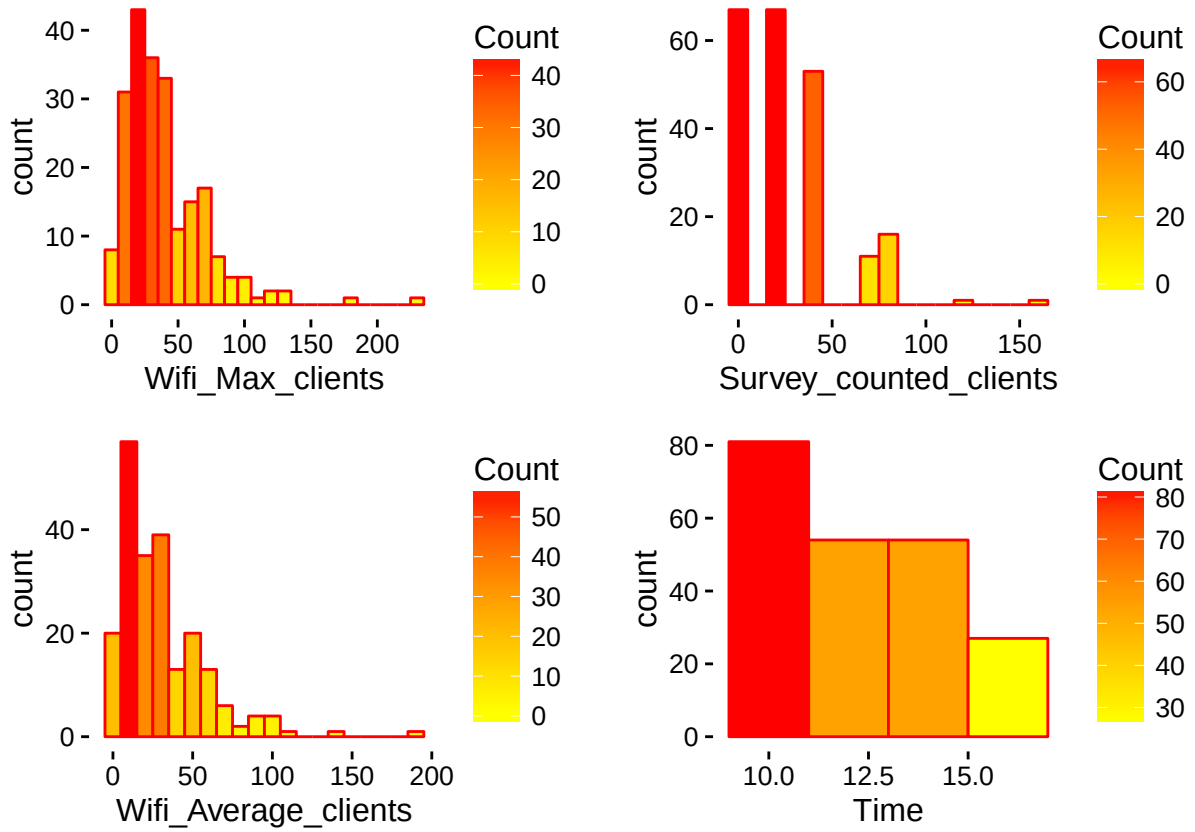
From this we could see that NaN values were not present in the data set. We could notice that the observations for the features Tutorials and Double_model were not even distributed across the 2 levels of the variables. In fact, only 6 observations were present for tutorial class and for double module class. Therefore, we decided to discard both the features, because they will be not informative for the analysis. Similarly for the feature class_went_ahead the majority of the lectures did occur and we decided to discard it. Furthermore, for the variables Wifi_Average_clients, Wifi_Max_clients and Survey_counted_clients it seems that there are few outliers, since the median is lower than the mean and the max values are far higher than the mean values. We will going to explore this issues with histogram and boxplots.

Exploratory graphs

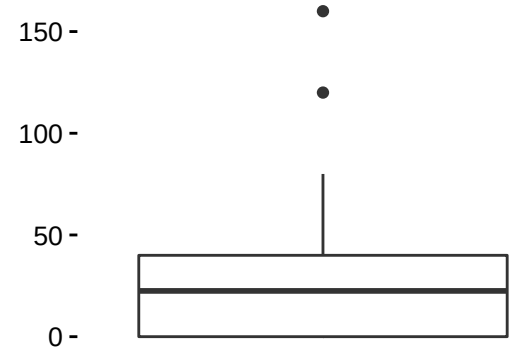
For exploring possible issues related with the continuous variables we plotted histograms and boxplots.

Histograms

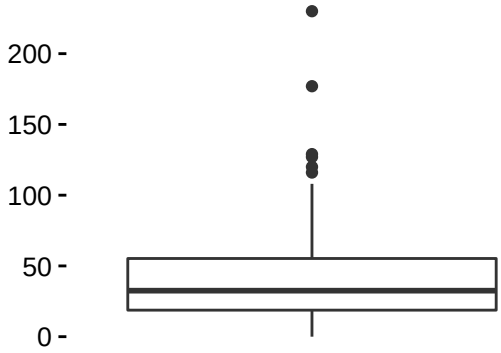
```
## Loading required package: grid
```



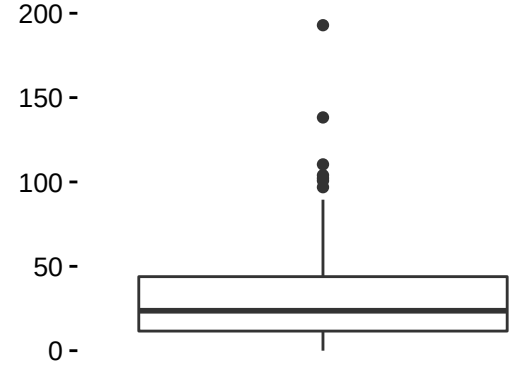
Form the histograms we could see that the distribution of the feature Wifi Maximum_client (i.e. the Maximum number of devices logged in one hour lecture) was skewed to the left, indicating that the in the majority of the lecture were counted no more than 40 people. Furthermore, we could see that there are potential outliers (values > 150). Similar pattern was observed for the feature Wifi_Average_clients. Different was the situation of the target feature, Survey_counted client, which showed a skewed distribution, but more scattered, similar to a Poisson distribution. This can cause a problem in running a linear regression and more likely we have have to run a generalise linear model with a Poisson distribution. This is not surprising, since we are dealing with count data (Zuur et al. 2009). Feature times had as well a skewed distribution, suggesting that the majority of the lectures were concentrating during the early morning and they were decreasing towards the afternoon.



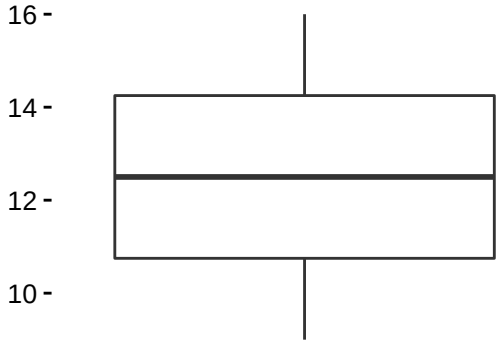
Counted clients with the survey



Maximum counted students with Wifi



Average Maximum counted students with

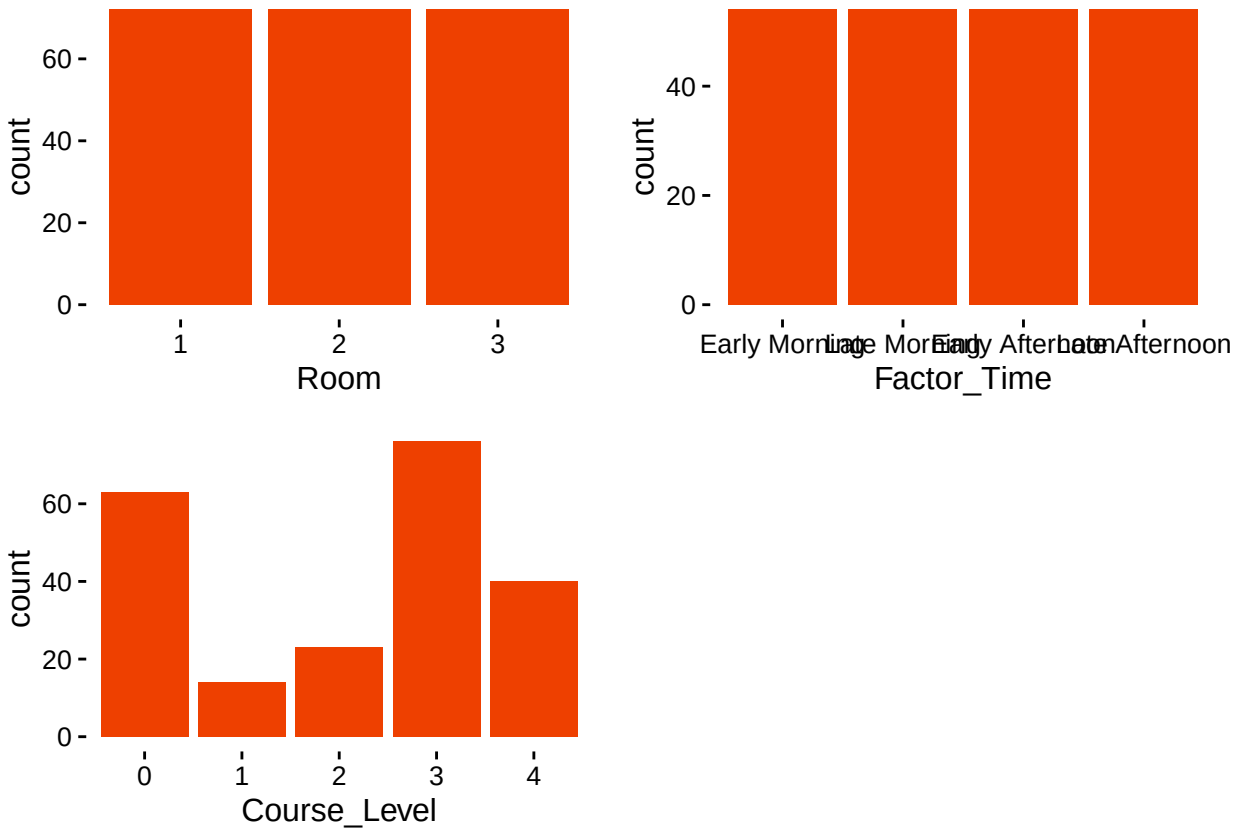


Maximum counted students

Boxplots

From the boxplots, all the trends observed in the histograms were confirmed.

For categorical variables we plot bar plot graphs.



Bar plots

From the barplots, we could see that observations were equally distributed across all the levels of the feature Room and Factor Time. On the contrary, there were more observations for non lectures and level 3 courses. No issues were detected for those features.

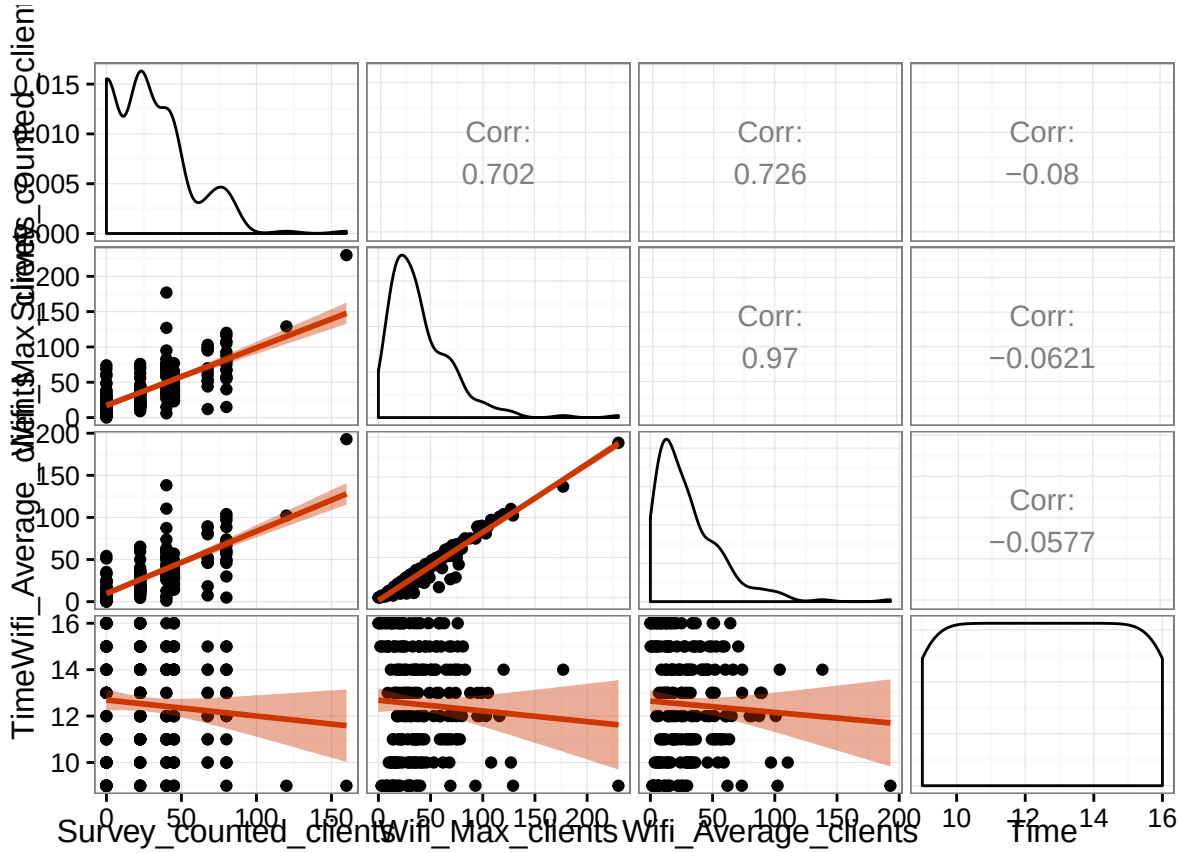
Features	Issues	Planned Solution
Room	None	None
Time	Distibution skewed to the left	To solve during analysis
Factor Time	None	None
Course level	None	None
Tutorial	Uneven representation of the level	Discarded from the analysis
Double Module	Uneven representation of the level	Discarded from the analysis
Class went ahead	Uneven representation of the level	Discarded from the analysis
Wifi Average clients	Distibution skewed to the left & outliers	To solve during analysis
Wifi Maximum clients	Distibution skewed to the left & outliers	To solve during analysis
Survey Counted clients	Distibution skewed to the left & outliers	To solve during analysis

Summary

FEATURES AFFECTING THE TARGET FEATURE

The next step of the analysis was to see which feature really affect the target feature for deciding which features we would include into the model.

Correlation matrix for continuous variables.



From the correlation matrix Survey counted clients seems to have a good correlation with Wifi Average counted clients and Maximum counted clients, therefore we are will try to run 2 models: one for exploring the relationship between Survey counted clients and Average counted clients and another for Survey counted clients and Maximum counted clients. However, from this graphs we can see that there is one point that is clearly two outliers. Therefore, we are going to run the analyses with and without them.

From the graphs we can see that Average counted clients and Maximum counted clients are highly correlated showing that both of them are not so different. Therefore we will not expect too much difference among the 2 models.

Time does not seems to be correlated with the target features Survey counted clients and it seems more categorical.